DOCUMENT RESUME

ED 305 376 TM 012 879

AUTHOR Suen, Hoi K.; And Others
TITLE Generalizability Assessment of Autocorrelated Direct Observation Data: The Applicability of the Tiao-Tan Method and Alternative.
PUB DATE Feb 88
NOTE 25p.; Paper presented at the Annual Meeting of the Eastern Educational Research Association (Miami Beach, FL, February 24-27, 1988).
PUB TYPE Reports - Evaluative/Feasibility (142) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Analysis of Variance; *Bayesian Statistics; Comparative Analysis; Equations (Mathematics); *Evaluation Methods; *Generalizability Theory; *Naturalistic Observation; Statistical Bias
IDENTIFIERS *Autocorrelation; Autoregressive Integrated Moving Averages; Box Jenkins Forecasting Model; Observer Reliability; Random Effects; Time Series Analysis; Variance (Statistical)

ABSTRACT

The applicability is explored of the Bayesian random-effect analysis of variance (ANOVA) model developed by G. C. Tiao and W. Y. Tan (1966) and a method suggested by H. K. Suen and P. S. Lee (1987) for the generalizability analysis of autocorrelated data. According to Tiao and Tan, if time series data could be described as a first-order autoregressive series with parameter "p" (rho), unbiased estimates of random error variance could be derived via a Bayesian process. Suen and Lee's two-step alternative procedure combines both Box-Jenkins time series analysis and a random-effect ANOVA process. The autocorrelated component of the data can be removed through the Box-Jenkins procedure, and then the residual or white-noise data can by analyzed via the ANOVA process to produce unbiased variance estimates. Theoretical advantages and limitations of the two approaches are outlined, focusing on autoregressive integrated moving averages. Three analyses of the methods are presented. Results from application of the methods to numerous data sets show that autocorrelation has a negligible or no effect on the systematic variance across observers. The Suen-Lee method is superior to the Tiao-Tan method in applications to generalizability assessment of observation data. Based on 28 behavioral observation time series, the Suen-Lee method seems to be applicable only when the relative systematic observer variance is small. A 29-item list of references and one data table are provided. (TJH)

**********
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
**********

ED305376

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
☑ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
Hoi K. Suen
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

# Generalizability Assessment of Autocorrelated Direct Observation Data: The Applicability of the Tiao-Tan Method and Alternative

Hoi K. Suen

University of Connecticut

Patrick S. C. Lee

Pennsylvania State University

and

Duncan K. H. Fong

Pennsylvania State University

Paper presentation at the Annual Meeting of the Eastern Educational Research Association, Miami Beach, Florida. February, 1988.

TM012879



BEST COPY AVAILABLE

## INTRODUCTION

The purpose of this paper is to explore the applicability of the Bayesian random-effect ANOVA model developed by Tiao and Tan (1966) and a method suggested by Suen and Lee (1987a, 1987b) for the generalizability analysis of autocorrelated data.

The generalizability theory of measurement (Cronbach et al., 1972, Brennan, 1983) has been suggested by many as particularly suited for the assessment of data acquired through the direct observation of behavior (e.g., Bakeman & Gottman, 1986; Berk, 1979; Cone, 1978; Hartmann, 1982; Kazdin, 1977; Mitchell, 1979), a method commonly used in classroom research, special education, and clinical and counseling psychology. A major obstacle to this application is the presence of autocorrelation often found among direct observation time-series data (cf. Gardner et al., 1982; Glass et al., 1975; Gottman & Glass, 1978; Hartmann et al., 1980; Jones et al., 1977, 1978). It is generally recognized that in the presence of correlated errors, estimates of expected mean squares (Box, 1954), expected variance components (Brennan, 1984), and subsequently reliability indices such as KR-20 (Maxwell, 1968) are biased, potentially leading to false indications of data reliability or a lack thereof.

The extent to which data acquired through the direct observation of behavior are autocorrelated is currently unclear and controversial. On the one hand, while authors such as Gardner et al. (1982), Glass et al., (1975); Gottman & Glass (1978), and Hartmann et al. (1980) suggested or deduced that many behavioral data are quite likely to be autocorrelated, Jones et al. (1977, 1978) produced empirical evidence that the majority of published behavioral data <u>are</u> in fact autocorrelated. This conclusion was severely challenged by

Huitema (1985), who reanalyzed Jones et al.'s data as well as other published data and concluded that the overwhelming majority of data acquired through direct observations of behavior are not autocorrelated. Huitema's conclusion of no autocorrelation was, in turn, challenged by others on methodological grounds (e.g., Suen, 1987; Suen & Ary, 1987; Sharpley & Alavosius, in press; Marascuillo & Busk, in press). Currently, the only conclusion one can draw is that it is inconclusive as to whether autocorrelation is a prevailing characteristic of observational data.

Although it is inconclusive as to the extent to which observational data are autocorrelated, everyday experiences suggest that at least some of the behavioral data are autocorrelated. When data are acquired by observing the occurrences of a behavior over time, it is reasonable to assume that, at least in some cases, the extent of occurrence of the behavior at a point in time is influenced by the extent of occurrence of the same behavior at previous points in time.

In the presence of autocorrelation, to produce unbiased estimates of variance components, which is an essential step in generalizability analysis (i.e., for both G and D studies), at least two methods have been developed. Tiao and Tan (1966) developed a Bayesian process for a one-way ANOVA which can be used when the nature of the autocorrelation can be described as a first-order autoregressive model (i.e., AR(1)). This method can theoretically be extended to an n-way ANOVA as well as higher order autoregressive models. Suen and Lee (1987a, 1987b), on the other hand, developed a two-step process by combining the Box-Jenkins (Box & Jenkins, 1976) autoregressive integrated moving average (ARIMA) modelling process with ANOVA. There are some theoretical advantages and disadvantages to each of these two methods.

However, possible applied problems associated with these two methods have not been explored. The purpose of this paper is to explore these possible problems.

## THE TIAO-TAN METHOD

Tiao and Tan (1966) suggested, if the data in a time series can be described as a 1st-order autoregressive series with parameter $\rho$, unbiased estimates of random error variance can be derived through a Bayesian process. For a k-observer by n-points of observation matrix, assuming a uniform prior distribution (i.e., no prior information), the posterior distribution of random error variance is reflected by the posterior distribution of $V=1/\sigma^2$, which is:

$$p(V|\rho,y) = S_1(\rho) f_\tau(k(m-1)) \frac{G_{\frac{1}{2}\phi(\rho)\tau}(\frac{1}{2}(k-1))}{H_{\phi(\rho)}(\frac{1}{2}(k-1),\frac{1}{2}k(m-1))}. \qquad (1)$$

where $y$ is the data vector of an observer, $G$ is the Gamma function at $\frac{1}{2}\phi(\rho)\tau$ with parameter $\frac{1}{2}(k-1)$, $m$ is $n-1$, and $S_1(\rho)$ is defined as:

$$S_1(\rho) = b_1 + c_1(\rho-\hat{\rho}_1)^2,$$

where $$\hat{\rho}_1 = \left\{\sum_{i=1}^{k}\sum_{j=2}^{m+1}(y_{ij}-\bar{y}_i^{(+)})(y_{i(j-1)}-\bar{y}_i^{(-)})\right\} \Big/ \left\{\sum_{i=1}^{k}\sum_{j=1}^{m}(y_{ij}-\bar{y}_i^{(-)})^2\right\},$$

$$b_1 = \sum_{i=1}^{k}\sum_{j=2}^{m+1}(y_{ij}-\bar{y}_i^{(+)})^2 - \hat{\rho}_1^2 \sum_{i=1}^{k}\sum_{j=1}^{m}(y_{ij}-\bar{y}_i^{(-)})^2,$$

$$\bar{y}_i(+) = \frac{1}{m} \sum_{j=2}^{m+1} y_{ij},$$

$$\bar{y}_i(-) = \frac{1}{m} \sum_{j=1}^{m} y_{ij},$$

and $$c_1 = \sum_{i=1}^{k} \sum_{j=1}^{m} (y_{ij} - \bar{y}_i(-))^2.$$

Additionally, for Eq. (1), $f_\tau(k(m-1))$ is the density at $\tau$ of a $\chi^2$ function with $k(m-1)$ degrees of freedom where $\tau = S_1(\rho)/\sigma^2$. The $\phi(\rho)$ value for the Gamma function in Eq. (1) is defined as $\phi(\rho) = S_2(\rho)/S_1(\rho)$, where

$$S_2(\rho) = b_2 + c_2(\rho - \hat{\rho}_2)^2,$$

$$\hat{\rho}_2 = \left\{ \sum_{i=1}^{k} (\bar{y}_i(+) - \bar{y}(+))(\bar{y}_i(-) - \bar{y}(-)) \right\} \Big/ \left\{ \sum_{i=1}^{k} (\bar{y}_i(-) - \bar{y}(-))^2 \right\}$$

$$b_2 = m \sum_{i=1}^{k} (\bar{y}_i(+) - \bar{y}(+))^2 - m\hat{\rho}_2^2 \sum_{i=1}^{k} (\bar{y}_i(-) - \bar{y}(-))^2,$$

$$c_2 = m \sum_{i=1}^{k} (\bar{y}_i(-) - \bar{y}(-))^2,$$

$$\bar{y}(+) = \frac{1}{k} \sum_{i=1}^{k} \bar{y}_i(+),$$

and $\bar{y}^{(-)} = \frac{1}{k} \sum_{i=1}^{k} \bar{y}_i^{(-)}$.

Finally, for Eq. (1):

$$H_{\phi(\rho)}(\tfrac{1}{2}(k-1),\tfrac{1}{2}k(m-1)) = I_{\phi(\rho)/(1+\phi(\rho))}(\tfrac{1}{2}(k-1),\tfrac{1}{2}k(m-1)),$$

with the right side of the equation being the beta function at $\phi(\rho)/(1+\phi(\rho))$ with parameters $\frac{1}{2}(k-1)$ and $\frac{1}{2}k(m-1)$. With the known posterior distribution of error variance as defined in Eq. (1), unbiased estimates of variance components can be obtained by identifying the modal maximum likelihood variance of the posterior distribution.

## THE SUEN-LEE ALTERNATIVE

Suen and Lee (1987a) proposed a two-step alternative procedure which combines both Box-Jenkins time series analysis (Box & Jenkins, 1976) and the random effect ANOVA process. First, the autocorrelated component of the data can be removed through the Box-Jenkins procedure and then the residual or white-noise data can be analyzed through the ANOVA process to produce unbiased variance estimates. For a T-time by K-observer single-facet crossed-design random effect generalizability analysis with serial dependency along T, the time dimension, the total variance can be decomposed as follows:

$$\sigma_x^2 = \sigma_t^2 - \omega_t + \sigma_k^2 + \sigma_{tk}^2 - \omega_{t*}, \qquad (2)$$

where $\omega_t$ and $\omega_{t*}$ are "moderating" effects on the T and the interaction variances due to autocorrelation. Both $\omega_t$ and $\omega_{t*}$ are negative in value because it was deduced that autocorrelation should lead to underestimates of the variance for T and the random error variance. This can be verified from the characteristic of t-ratios. In the presence of autocorrelation, it is commonly known (e.g., McDowall et al., 1978; Sharpley & Alavosius, in press) that the t-ratio is spuriously inflated. It is also known that a mean is a deterministic statistic and is not influenced by the presence of autocorrelation. Hence, the only explanation for the inflation of t is that the standard error is underestimated. In other words, the variance is underestimated.

Suen & Lee suggested that unbiased estimates of variance components can be obtained by applying the Box-Jenkins procedure (Box & Jenkins, 1976) on the vector of mean scores across observers to identify the best ARIMA(p,d,q) model. The general ARIMA model for the mean, $M_i$, at the ith point in time is:

$$M_i = E_i + \phi_1 E_{i-1} + \phi_2 E_{i-2} + \ldots + \phi_p E_{i-p} + \theta_1 M_{i-1} + \theta_2 M_{i-2} + \ldots + \theta_q M_{i-q}, \qquad (3)$$

where $E_i$ is the random or white noise component at time i, $\phi_p$ is the pth order autoregressive parameter, and $\theta_q$ is the qth order moving average parameter. The identified model can then be applied to the data of each observer individually to "filter" autocorrelation from the data. Specifically, the score from the ith observation of the jth observer can be "filtered" through:

$$X'_{ij} = X_{ij} - \hat{\phi}_1 X'_{(i-1)j} - \ldots - \hat{\phi}_p X'_{(i-p)j} - \hat{\theta}_1 X_{(i-1)j} - \ldots - \hat{\theta}_q X_{(i-q)j}, \qquad (4)$$

where $X'_{ij}$ is the "filtered" and $X_{ij}$ is the raw score at the ith point in time for the jth observer.

An ANOVA of the "filtered" data matrix with elements $X'_{ij}$ would produce unbiased variance estimates. The ω values in Eq. (2) can then be obtained by contrasting the differences of the results of ANOVA's on the "filtered" and "unfiltered" data. A problem with Eq. (4) is that $X_{1j}$ has to be assumed to contain only a random component since $X_{(1-p)j}$ is not available. This is counter-intuitive. To resolve this problem, Suen & Lee (1987b) later suggested the use of backforecasting techniques (Box & Jenkins, 1976).

## THEORETICAL ADVANTAGES AND LIMITATIONS OF THE TWO METHODS

For the purpose of generalizability assessment, a fundamental problem with the Suen-Lee proposal is that, through the ARIMA process, the random errors of variance estimates (i.e., variances of variance estimates, $VAR(\hat{\sigma}_x^2)$, $VAR(\hat{\sigma}_k^2)$, & $VAR(\hat{\sigma}_{tk}^2)$) are expected to be inflated by the error of specification of the ARIMA model. Specifically, because the ARIMA analysis is performed on the mean score series, not the data streams of individual observers, and yet individual observer series were filtered based on the model for the mean series, errors in specifying the model for the mean series as the representative series for all observers will compound the random error of the variance estimates. Suen & Lee's argument was that if the inflation of these variances of variance estimates is small, the process is effective. If they prove to be large, systematic errors are replaced by random errors. The exact amount of inflation, however, is unknown. The major advantage of the Suen-Lee

method is that it can be applied to any autocorrelated data regardless of the nature of the autocorrelation. That is, the Suen-Lee method can be applied to any ARIMA model as well as cyclical SARIMA model.

The Tiao-Tan method, on the other hand, offers several advantages over the Suen-Lee procedure. First, instead of being an ARIMA method, the Tiao-Tan method is a random effects model and is hence consistent with the generalizability theory. Second, the variances of the final variance components can be derived. Finally, the inflation of these variances of variances can be expected to be smaller than those with the Suen-Lee method.

On the other hand, the Tiao-Tan method contains some limitations. First, as a Bayesian process, some assumptions have to be made regarding the form of the prior distribution. Tiao and Tan pointed out that different likelihood functions, and subsequently posterior distributions of variance estimates, will result if a normal prior distribution rather than a uniform prior distribution is assumed. Second, the method is appropriate only if the data form a 1st-order autoregressive series. Although Tiao and Tan suggested that their method can be extended to higher-order autoregressive series, it is inappropriate for moving average series ARIMA(0,0,q), mixed series ARIMA(p,0,q), and seasonal or cyclical series SARIMA(p,d,q). Finally, and perhaps most seriously from an applied perspective, the Tiao-Tan method requires derivations of complex Gamma and Beta functions. The complexity is particularly severe with the Gamma function. Since computer softwares are not available, the use of the Tiao-Tan method also depends on the ability to develop and compile a complex computer algorithm.

Both the Tiao-Tan method and the Suen-Lee method are appropriate only for a single-facet random-effect fully-crossed-design generalizability assessment.

While the Suen-Lee method is specifically designed for the random model involving a two-way ANOVA, the Tiao-Tan method is designed for a one-way ANOVA.

## UNANSWERED QUESTIONS

With the limitations of the two methods, three major questions remain as to the applicability of the Tiao-Tan method and the Suen-Lee method for generalizability assessment. First, the Tiao-Tan method appears to offer some theoretical advantages over the Suen-Lee method, even considering the complex Gamma function for the former. Since the Tiao-Tan method is applicable only to ARIMA(1,0,0) series, the critical question is how prevalent are ARIMA(1,0,0) series in behavioral observation data? If ARIMA(1,0,0) series are prevalent, Tiao-Tan appears to be the method of choice for generalizability assessments of autocorrelated behavioral observation data. Considerations should then be given to the development of appropriate software for the use of the method.

If ARIMA(1,0,0) series is rare, the Suen-Lee method would be more useful since it applies to all forms of ARIMA(p,d,q) as well as SARIMA(p,d,q) models. The question then is whether the Suen-Lee method can be refined so that the expected inflation of the variances of variances due to misrepresentation by the mean series can be minimized? If it can be refined, what applied problems may arise with the refined approach?

To answer this series of questions, three analyses were conducted as below.

## FIRST ANALYSIS

Given the current uncertainty regarding the extent to which behavioral data are autocorrelated (see earlier discussion), existing literature will not offer an answer to the first question of whether ARIMA(1,0,0) series are prevalent. Nor would a secondary analysis of published data offer an answer since most observational data in the literature contain extremely short series (cf. Huitema, 1985; Suen & Ary, 1987). Hence, a sample of primary data is used to assess the extent to which observational data contain ARIMA(1,0,0) series.

### SAMPLE DATA

The raw observational data in Ary et al.'s (1986) study were used. Ary et al. employed 7 observers to record 4 behaviors of a child at 5-second intervals for 1 hour, generating a total of 720 observations per observer per behavior, or a total of 20,160 observation data points. Each data point is dichotomous (occur vs. not occur). For our purpose, each of the 720-point dichotomous series was collapsed into a 144-point time series of prevalence scores (or proportion of time behavior occurs) by transforming dichotomous scores in every 5 5-second intervals (representing 25 second real time) into a prevalence proportion (or p or mean score). Subsequently, a total of 28 144-point time series representing the prevalence of 4 behaviors of a child as recorded by 7 independent observers were obtained.

### DATA ANALYSIS

To investigate the extent to which ARIMA(1,0,0) series exists in behavioral data, the 28 time-series were submitted to 28 separate ARIMA

analyses. Specifically, analyses were performed through the ARIMA routine in MINITAB. In addition to estimating ARIMA parameters through the judgments of autocorrelation function (ACF) and partial autocorrelation function (PACF) correlograms, the t-ratios of the estimated ARIMA parameters were examined to determine the significance of the parameter estimates.

## RESULTS AND CONCLUSION

Of the 28 time series, none yielded a statistically significant 1st-order autoregressive (AR1) parameter. One of the 28 series yielded statistically significant ARIMA(1,0,1) parameters ($\hat{\phi}_1$=.774, t=3.30, p<.05; $\hat{\theta}_1$=.655, t=2.36, p<.05) and four yielded statistically significant seasonal autoregressive (i.e., SARIMA(9,0,0)) parameters, with cycles of 9 data points ($\hat{\phi}$ = .217, .376, .363, & .266; t = 2.44, 4.37, 4.08, & 3.03; p<.05). In other words, of a sample of 28 behavioral observation time series, 5 (i.e., 18%) showed some form of autocorrelation. However, none of these autocorrelated data can be described as an ARIMA(1,0,0) series.

To the extent that the limited sample of 28 data sets can represent typical behavioral observation data, a conclusion can be drawn that autocorrelation does exist in a substantial portion (about 1 out of 5) of observational data. At the same time, an ARIMA(1,0,0) series is rare (none at all in this sample). Hence, the Tiao-Tan method may have only very limited utility for behavioral observation data.

## SECOND ANALYSIS

The lack of ARIMA(1,0,0) series in behavioral observation data is not a

problem with the Suen-Lee method, which can be applied to all ARIMA and SARIMA series. One way to minimize the expected inflation in the variances of variance estimates is to limit the application of the Suen-Lee method to situations in which the ARIMA or SARIMA models and parameters for different observers of the same behavior are similar. In such a case, using the ARIMA or SARIMA model of the mean series as the representative model for the data from all observers will contain a minimal amount of misrepresentation. When this occurs, the inflation of the variances of variance estimates can be expected to be minimal or negligible.

For the purpose of the present analysis, two criteria were used to judge if the K series from K observers of the same behavior are sufficiently similar to be represented by the ARIMA parameters of the mean series. First, the series for individual observers need to fit the same ARIMA(p,d,q) or SARIMA(p,d,q) model. Second, the values of the K parameter estimates (i.e., $\hat{\phi}$'s and $\hat{\theta}$'s as appropriate) need to be within one standard error of estimate of one another (i.e., middle 68% interval). The question is how often can we expect different observers recording the same autocorrelated behavior to yield data that meet these two criteria?

## DATA ANALYSIS AND RESULTS

The results of the ARIMA analysis of the 5 sets of autocorrelated data from Ary et al.'s study were examined pairwise to find the frequency with which the two criteria are met. Of the 5 sets of autocorrelated data, 4 met the two criteria of similarity. Specifically, the 3rd and 7th and the 4th and 6th observers of the "questioning" behavior produced two pairs of data sets that met the two criteria. (Note: had a less stringent criteria of 95% confidence

interval instead of the 68% interval been used, all four sets were within 2 standard errors of one another.) All 4 data sets can be described as SARIMA(9,0,0) models with cycles of 9 data points. The estimated SAR(9) parameters are .217 (68%CI: .128-.306) for the 3rd observer, .266 (68%CI: .178-.354) for the 7th observer, and .376 (68%CI: .290-.462) for the 4th observer, and .363 (68%CI: .274-.452) for the 6th observer. In other words, while all SAR(9) parameter estimates were well within the 95%CI of one another, the 3rd and 7th observers and the 4th and 6th observers formed distinct pairs of particularly similar parameters. Since 4 out of 5 autocorrelated series met the two conditions specified above for the application of the Suen-Lee method with negligible inflation of variances of variance estimates, one may conclude that restricting the use of the Suen-Lee method to situations in which the two criteria are met may not seriously lessen its applicability.

## THIRD ANALYSIS

To investigate whether unexpected problems may arise in the application of the Suen-Lee method when the two criteria of similarity are met, the data from the 3rd and 7th and those from the 4th and 6th observers were submitted to two separate analyses through the Suen-Lee method. To determine the efficacy of the Suen-Lee method used under the restriction of similar ARIMA or SARIMA parameters, the variance estimates before and after filtering were compared. Based on the model in Eq. (2), the variance estimates of the filtered data should be larger than their unfiltered data counterparts. In other words, if any the filtered variance estimates is smaller than its unfiltered counterpart, either the process is not effective or the model in Eq. (2) is misspecified.

DATA ANALYSIS AND RESULTS

The Suen-Lee method was applied to the two pairs of data sets and the backforecasting method was used to filter autocorrelation from the data. For the purpose of filtering, the SARIMA(9,0,0) parameters of the mean series were used. Specifically, the $\phi$ estimate for the mean series between the 3rd and 7th observers was .2805 (std. dev.=.0885, t=3.17, p<.05) and it was .4036 (std. dev.=.0862, t=4.68, p<.05) for the mean series between the 4th and 6th observers. The filtered data were then submitted to a random-effect two-way ANOVA process in the usual fashion. The variance estimates were obtained through the VARCOMP procedure in SAS, as recommended by Bell (1985). Table 1 presents the subsequent filtered and unfiltered variance estimates and associated generalizability coefficients.

---
Insert Table 1 about here
---

For both observer pairs, the interaction (or random error) variance estimates for the filtered data were smaller than those of the unfiltered data, as hypothesized. For the 3rd-7th observer pair, the interaction variance from the unfiltered data shows an underestimate of only 1.3% when compared against the filtered variance estimate. For the 4th-6th observer pair, however, the interaction variance was underestimated by 10%. Since the SARIMA(9,0,0) parameter estimate for the 3rd-7th pair was also smaller than that for the 4th-6th pair (i.e., .2805 vs. .4036), it provides limited support to the belief that the biased effects of autocorrelation on interaction variance estimate is a direct function of the magnitude of the autoregressive parameter (e.g., Tiao

& Tan, 1966).

The systematic observer error variance (i.e., $\hat{\sigma}_k^2$) estimates remained unchanged before and after filtering for both observer pairs. In all cases, the relative magnitudes of the estimates were extremely low (i.e., .00001 and .0000). The fact that observer variance remains unchanged is consistent with the Suen-Lee model specified in Eq. (2). With the wisdom of hindsight, however, the extremely small variance across observers should also have been predicted because of the criteria used. Given the criteria that the Suen-Lee method be applied only when the ARIMA or SARIMA parameters are similar, the systematic components of the data from different observers are by definition quite similar and that their data differ only in random or white-noise fluctuations. Hence, the systematic variance across observers are small. The implication for this small variance across observers is that subsequent estimates of criterion-referenced generalizability coefficients would be very similar to their norm-referenced counterparts. In any event, the findings confirm the hypothesis that autocorrelation will have negligible or no effect on the systematic variance across observers.

The estimations of systematic variance across time (i.e., $\hat{\sigma}_t^2$) provided mixed results. For the 3rd-7th observer pair, the filtered $\hat{\sigma}_t^2$ estimate was indeed larger than the unfiltered $\hat{\sigma}_t^2$, as hypothesized. However, for the 4th-6th observer pair, the opposite result was obtained. This suggests that either the filtering process has failed to correct for the effect of autocorrelation on $\sigma_t^2$ or the model as described in Eq. (2) was misspecified. Based on the fact that previous applications of the method (i.e., Suen and Lee, 1987a, 1987b) as well as estimates for other variance components in the present study have produced results as predicted, it is most

likely that the effect on $\sigma_t^2$ in Eq. (2) (i.e., $-\omega_t$) was misspecified and that, in fact, $\omega_t$ can be either positive or negative in value.

More careful logical deduction also supports the notion that $\omega_t$ can be either positive or negative. The hypothesis that $\omega_t$ is negative was based on the known characteristic of the t-ratio in the presence of autocorrelation. However, the underestimation of variance in the t-ratio reflects only the effect of autocorrelation on <u>random</u> variance, not <u>systematic</u> variance. Since variance across time is a systematic variance, the effect of autocorrelation on this variance cannot be predicted from the t-ratio deduction. Hence, the more appropriate model is:

$$\sigma_x^2 = \sigma_t^2+\omega_t+\sigma_k^2+\sigma_{tk}^2+\omega_{t*},$$

where $\omega_t$ can take on a positive or negative value, and $\omega_{t*}$ will always be negative.

The differences between the filtered and unfiltered common dependability coefficients (i.e., the norm-referenced generalizability coefficient, $\hat{\rho}^2$; the signal/noise ratio, $\Psi(g)$; the criterion-referenced generalizability coefficient, $\Phi$, and the norm-referenced Cronbach alpha) were practically inconsequential for the 3rd-7th pair, for which the differences were noticeable only at the 3rd decimal place. The differences were, however, more pronounced for the 4th-6th observer pair. This may suggest again that the effect of autocorrelation on the results of a generalizability analysis is a direct function of the magnitude of the ARIMA parameter.

## SUMMARY AND DISCUSSION

Based on a sample of 28 behavioral time series, it was found in the first analysis that 5 (i.e., 18%) contained some form of autocorrelation. However, none of these could be described as an ARIMA(1,0,0)) model. When combined with the necessity to calculate a Gamma function within the Tiao-Tan method, this suggests that the Tiao-Tan method may have severely limited applied value for the generalizability assessment of behavioral observation data.

The Suen-Lee method appears to be a viable alternative provided that the inflation of the variances of variance estimates can be minimized. One method to attain this is to use the Suen-Lee method only when the ARIMA parameters for the data from different observers are sufficiently similar. Using the criteria of identical ARIMA or SARIMA models across observers and the parameter estimates being within one standard error of one another, it was found that 4 out of 5 autocorrelated series could be considered sufficiently similar such that the application of the Suen-Lee method will lead to relatively stable variance estimates.

An implied result of limiting the application of the Suen-Lee method to situations in which the ARIMA parameters across observers are similar is that the Suen-Lee method is applicable only to situations in which the relative systematic observer variance is small. For these situations, the differences between estimates for a norm-referenced vs. a criterion-referenced interpretation becomes negligible.

It was also found that the effect of autocorrelation on the systematic variance across time is not always negative as previously hypothesized. To summarize, in the presence of autocorrelation, the systematic variance across

time may be over- or under-estimated. Autocorrelation along time will have no effect on the systematic variance across observers. The interaction or random error variance will always be underestimated in the presence of autocorrelation.

A limitation of the above analyses is that they were based on only 28 behavioral observation time series, despite the fact that 20,160 original empirical data points were involved. Given that there were only 28 time series, the results may not be representative of all behavioral time series. Hence, the conclusions of these analyses can only be considered tentative.

To investigate the representativeness of the findings of the analyses in this paper, two other sets of behavioral data are currently being analyzed. Specifically, 25 time series in Brulle and Repp's (1984) study representing 5 classroom behaviors of a child over 5 days with a total of 44,000 observation points and 32 time series in Brusca's (1986) study representing 2 stereotypic behaviors of an autistic child over 16 observation sessions are being analyzed. Preliminary evidence suggests that the existence of ARIMA(1,0,0) series in these data may also be rare and that seasonal models are common, particularly among stereotypic behaviors.

## References

Ary, D., Van Acker, R. M., & Karsh, K. L. (1986). Comparing behavior duration reliabilities of momentary time-sampling and continuous electronic data gathering. Paper presented at the Annual Meeting of the American Educational Research Association, San Francisco. April.

Bakeman, R., & Gottman, J. M. (1986). Observing interaction: An introduction to sequential analysis. London: Cambridge University Press.

Berk, R. A. (1979). Generalizability of behavioral observations: A clarification of interobserver agreement and interobserver reliability. American Journal of Mental Deficiency, 83, 460-472.

Bell, J. F. (1985). Generalizability theory: The software problem. Journal of Educational Statistics, 10 (1), 19-29.

Box, G. E. P. (1954). Some theorems on quadratic forms applied in the study of analysis of variance problems, II: Effects of inequality of variance and of correlation between errors in the two-way classification. Annals of Mathematical Statistics, 29, 885-891.

Box, G. E. P., & Jenkins, G. M. (1976). Time series analysis: Forecasting and control (revised ed.). San Francisco: Holden-Day.

Brennan, R. L. (1983). Elements of generalizability theory. Iowa City: ACT Publications.

Brennan, R. L. (1984). Some statistical issues in generalizability theory. Invited symposium paper presented at the annual meeting of the American Educational Research Association, New Orleans. April.

Brulle, A. R., & Repp, A. C. (1984). An investigation of the accuracy of momentary time sampling procedures with time series data. British Journal of Psychology, 75, 481-485.

Cone, J. D. (1978). The relevance of reliability and validity for behavior assessment. Behavior Therapy, 8, 411-426.

Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. (1972). The dependability of behavioral measurements: Theory of generalizability for scores and profiles. New York: John Wiley & Sons.

Gardner, W., Hartmann, D. P., & Mitchell, C. (1982). The effects of serial dependency on the use of chi-square for analyzing sequential data. Behavioral Assessment, 4, 75-82.

Glass, G. V., Willson, V. L., & Gottman, J. M. (1975). Design and analysis of time-series experiments. Boulder, CO: Univ. of Colorado Press.

Gottman, J. M., & Glass, G. V. (1978). Analysis of interrupted time-series experiments. In T. R. Kratochwill (Ed.), Single subject research: Strategies for evaluating change. New York: Academic Press.

Hartmann, D. P. (1982). Assessing the dependability of observational data. In D. P. Hartmann (Ed.), Using observers to study behavior. San Francisco: Jossey-Bass.

Hartmann, D. P., Gottman, J. M., Jones, R. R., Gardner, W., Kazdin, A. E., & Vaught, R. S. (1980). Interrupted time-series analysis and its application to behavioral data. Journal of Applied Behavior Analysis, 13, 543-559.

Huitema, B. E. (1985). Autocorrelation in applied behavior analysis: A myth. Behavioral Assessment, 7, 107-118.

Jones, R. R., Vaught, R. S., & Weinrott, M. (1977). Time-series analysis in operant research. Journal of Applied Behavior Analysis, 10, 151-166.

Jones, R. R., Weinrott, M., & Vaught, R. S. (1978). Effects of serial dependency on the agreement between visusal and statistical inference. Journal of Applied Behavior Analysis, 11, 277-283.

Kazdin, A. E. (1977). Artifacts, bias, and complexity of assessment: The ABC's of reliability. Journal of Applied Behavior Analysis, 10, 141-150.

Maxwell, A. E. (1968). The effect of correlated errors on restimates of reliability coefficients. Educational and Psychological Measurement, 28, 803-811.

McDowall, D., McCleary, R., Meidinger, E. E., & Hay, R. A., Jr. (1980). Interrupted time series analysis. In J. L. Sullivan (Ed.), Quantitative applications in the social sciences. Beverly Hills, CA: Sage.

Mitchell, S. K. (1979). Interobserver agreement, reliability, and generalizability of data collected in observational studies. Psychological Bulletin, 86, 376-390.

Sharpley, C. F., & Alavosius, M. P. (In press). Autocorrelation in behavioral data: Some facts. Behavioral Assessment.

Suen, H. K. (1987). On the epistemology of autocorrelation in applied behavior analysis. Behavioral Assessment, 9.

Suen, H. K., & Ary, D. (1987). Autocorrelation in applied behavior analysis: Myth or reality? Behavioral Assessment, 9.

Suen, H. K., & Lee, P. S. C. (1987a). An aggregate-segregate approach to the generalizability assessment of data of correlated errors. Paper presented at the annual meeting of the American Education Research Association, Washington, DC. April.

Suen, H. K., & Lee, P. S. C. (1987b). The generalizability assessment of autocorrelated data via Box-Jenkins backforecasting. Paper presented at the annual meeting of the Midwestern Educational Research Association, Chicago. October.

Tiao, G. C., & Tan, W. Y. (1966). Bayesian analysis of random effect models in the analysis of variance, Part II: Effects of autocorrelated error. Biometrika, 53, K-477-495.

Table 1

VARIANCE ESTIMATES OF UNFILTERED AUTOCORRELATED BEHAVIORAL DATA AND THE SAME DATA FILTERED THROUGH THE SUEN-LEE METHOD

| Observers | Parameter | Unfiltered Estimate | Filtered Estimate |
|---|---|---|---|
| 3rd & 7th | $\hat{\sigma}_t^2$ | .00749 | .00787 |
| | $\hat{\sigma}_k^2$ | .00001 | .00001 |
| | $\hat{\sigma}_{tk}^2$ | .00235 | .00238 |
| | $\hat{\rho}^2$ | .76112 | .76780 |
| | $\Psi(g)$ | 3.18723 | 3.30672 |
| | $\bar{\xi}$ | .76041 | .76706 |
| | Cronbach $\alpha$ | .86440 | .86865 |
| 4th & 6th | $\hat{\sigma}_t^2$ | .00592 | .00530 |
| | $\hat{\sigma}_k^2$ | .00000 | .00000 |
| | $\hat{\sigma}_{tk}^2$ | .00151 | .00167 |
| | $\hat{\rho}^2$ | .79677 | .76040 |
| | $\Psi(g)$ | 3.92053 | 3.17365 |
| | $\bar{\xi}$ | .79677 | .76065 |
| | Cronbach $\alpha$ | .88689 | .86390 |